AMERICAN UNIVERSITY OF BEIRUT

AUTOMATED DETECTION OF WOMEN DEHUMANIZATION IN ENGLISH TEXT

by
MAHA ABDUL JABBAR WISS

A thesis
submitted in partial fulfillment of the requirements
for the degree of Master of Science in Business Analytics
to the Suliman S. Olayan School of Business
at the American University of Beirut

Beirut, Lebanon
September 2022

AMERICAN UNIVERSITY OF BEIRUT

AUTOMATED DETECTION OF WOMEN DEHUMANIZATION IN ENGLISH TEXT

by

MAHA ABDUL JABBAR WISS

Approved by:

Dr. Wael Khreich, Assistant Professor — Advisor
Suliman S. Olayan School of Business

Dr. Wissam Sammouri, Assistant Professor of Practice — Member of Committee
Suliman S. Olayan School of Business

Dr. Sirine Taleb, Lecturer — Member of Committee
Suliman S. Olayan School of Business

Date of thesis defense: September 13, 2022

AMERICAN UNIVERSITY OF BEIRUT

THESIS RELEASE FORM

Student Name: Wiss Maha Abdul Jabbar

I authorize the American University of Beirut, to: (a) reproduce hard or electronic copies of my thesis; (b) include such copies in the archives and digital repositories of the University; and (c) make freely available such copies to third parties for research or educational purposes:

☐ As of the date of submission

☒ One year from the date of submission of my thesis.

☐ Two years from the date of submission of my thesis.

☐ Three years from the date of submission of my thesis.

Maha Wiss  15/09/2022

Signature Date

ACKNOWLEDGEMENTS

I would like to express my deepest appreciation to my advisor Dr. Wael Khreich for his guidance, for his continues support to stimulate the best we have, and for his invaluable patience and feedback throughout this study.
I would also extend my sincere thanks to Dr. Wissam Sammouri for his generous support and unforgettable advice.
Additionally, words cannot express my gratitude to MEPI-TLG team for generously giving me the opportunity to continue my master's degree and for funding all of my needs.
Lastly, I could not have undertaken this journey without the help and motivation of my partner Eng. Moatassim Alibrahim. His belief in me has kept my spirits high during this process.
This accomplishment would not have been possible without them. Thank you.

ABSTRACT

OF THE THESIS OF

Maha Abdul Jabbar Wiss for Master of Science in Business Analytics
Major: Business Analytics

Title: Automated Detection of Women Dehumanization in English Text

Animals, objects, food, plants, and other non-human terms are commonly used as a source of metaphors to describe females, in formal and slang language. Comparing women to non-human items not only reflects cultural views that might conceptualize women as subordinates or in a lower position than humans, yet it conveys this degradation to the listeners. Moreover, the dehumanizing representation of females in the language normalizes the derogation and, even, encourages sexism and aggressiveness against women. Although dehumanization has been a popular research topic for decades, according to our knowledge no studies have linked the women dehumanizing language to the machine learning field. Therefore, we introduce our research work as one of the first attempts to create a tool for the automated detection of the dehumanizing depiction of females in English texts. We, also, present the first labeled dataset on the charted topic, which is used for training supervised machine learning algorithms to build an accurate classification model. The importance of this work is that it accomplishes the first step toward mitigating dehumanizing language against females.

TABLE OF CONTENTS

ILLUSTRATIONS

Figure

TABLES

Table

ABBREVIATIONS

NLP	Natural Language Processing
UH	Uniquely Human
HN	Human Nature
IAT	Implicit Association Test
ASI	Ambivalent Sexism Inventory
LSH	Likelihood to Sexually Harass
ASA	Attraction to Sexual Aggression
RBA	Rape Behavioral Analogue
IRMA	Illinois Rape Myth Acceptance
HMM	Hidden Markov Model
CRF	Conditional Random Fields
NPOV	Neutral Point of View
SAGM	Sparse Additive Generative Models
LR	Logistic Regression
SVM	Support Vector Machine
Opt. SVM	Optimized Support Vector Machine
NB	Naïve Bayes
MNB	Multinomial Naïve Bayes
RF	Random Forest
XGB	Extreme Gradient Boosting
TF-IDF	Term Frequency Inverse Document Frequency
BoW	Bag of Words
TP	True Positives
FP	False Positives
TN	True Negatives
FN	False Negatives
TPR	True Positive Rate
FPR	False Positive Rate
ROC	Receiver Operating Characteristic
AUC	Area Under Curve
Raw	Raw Data
CE	Expanding Contractions
LC	Lowercasing
PR	Punctuation Removal
SW	Stop words Removal
L	Lemmatization
S	Stemming

CHAPTER 1

INTRODUCTION

Discrimination and bias against specific gender, race, religion, color, etc. have been widespread since ever. These biases take their forms from human culture, inherited beliefs, past experiences, religious opinions, political desires, and other sources. The reflection of human biases can be very explicit and direct, such as aggressive behavior or against a group of people. It can, also, be reflected indirectly through extremist thoughts and even semantically in human language.

Language is one of the most powerful instruments in committing and reproducing sexism and gender bias (Menegatti & Rubini, 2017). The reason might be that gender stereotypes about womanhood (e.g., nice, and caring) and manhood (e.g. self-assured, and agentic) are reflected and reproduced through the selection of lexical items in our daily communication (Cuddy et al., 2004). Perpetrating sexism through language can convey these stereotypes to the recipient's cognition and formulate actual discrimination (Menegatti & Rubini, 2017).

Sexism in language takes many forms and has several types. Recently, a comprehensive taxonomy that categorizes different types of gender bias in the text was proposed by Doughman et al., (2021). This taxonomy includes gender bias in semantics, which means hiding sexism and biases in the semantic meaning of words (Umera-Okeke, 2012). In other words, gender bias can be hidden implicitly behind jokes, old sayings, proverbs, and metaphorical and comparative expressions (Umera-Okeke, 2012). Examples include setting comparisons by using non-human items to describe humans, which represents a sense of dehumanization.

For decades, dehumanization has been one of the popular concepts in literature. Generally, dehumanization is defined as the denial of humanness (Haslam, 2006). To illustrate, humanness can be defined either by disclaiming characteristics that distinguish humans from animals (e.g., civility, logic, maturity), or by the denial of traits that are standard to humans (e.g., cognition, emotions, individuality). Therefore, comparing human to non-human items could be dehumanizing as these comparative expressions highlight some features and mask others. Particularly, comparing women to non-human items using metaphors, similes, or any other comparative structure, to slander and even sometimes to praise, could carry devaluating connotations toward described females. This derogation represents a kind of women dehumanization. (See examples in Table 1.1)

A common strategy used to study the sexist and dehumanizing depiction of women in the text is to analyze the metaphorical and comparative expressions used in describing females. This analysis involves a manual extraction of phrases describing women from different texts such as women's magazines, news, etc. Then, dividing non-human terms and metaphors, that are used as source-domain, into taxonomies either based on the similarity (e.g., animals, food, objects, etc.) or conceptually, according to the promoted concept of womanhood (e.g., sensitivity, weakness, reproductivity, motherhood, wifehood, talkative, power, resilience, sexuality, evil, and others). After the categorization, studies attempt to identify the potential semantic meanings and connotations of each category, which leads to a better investigation of the negative impact of this dehumanizing language.

Many works in the literature demonstrate that depicting females using food, objects, animals, and other non-human items belittles and trivializes women to their

sexuality (Baider & Gesuato, 2003; Hines, 1999; Kang, Hye-Min; Shaydullina, 2015). This derogation can be brought to the conscious level, and eventually, can negatively affect the human conceptual system. More importantly, experiments prove that presenting females using animalistic metaphors can influence actions by increasing men's rape myth acceptance, rape proclivity, and sexual aggressiveness against women (Bock & Burkley, 2019; Morris et al., 2018; Rudman & Mescher, 2012; Tipler & Ruscher, 2017). This means that the insistence on using this language might put more females at risk of harassment of any form.

Considering the deep harm that dehumanizing language causes to the individual consciousness and society, it turns out to be even more problematic that we lack existing solutions. According to our knowledge, no existing tools or systems that could help in the automated detection or mitigation of women dehumanization in text. This represents the gap in the literature and the actual need for creating a tool that has some power in identifying and detecting women dehumanizing phrases. Subsequently, once these statements are accurately isolated, mitigation becomes possible.

Working on the semantic level of the words is challenging even for humans who might be themselves influenced by rooted stereotypes. Below are some examples that clarify how subtle the difference is between sentences with the same key terms and pattern "women are …roses …", however, they carry different semantics (Table 1.1). Identifying dehumanization in these sentences is a complex process that requires human reasoning.

Given the difficulty of working on the semantics and detecting women dehumanization from text, even for humans. In addition to lacking automated tools that can be used for this purpose. Our research aims to illuminate this uncharted area by

building a bridge between linguistic and machine learning domains by creating a system that automates the detection of women dehumanizing depiction in English. For the system to be powerful and accurate in classifying sentences and detecting the dehumanization of females from English texts, it requires a labeled and comprehensive dataset, which is, also, one of the major gaps in this field. Proposing a well-built dataset will be also a key contribution to this study. To make this dataset representative and useful for the community, it will be carefully collected and validated through annotators' voting. Which will, consequently, facilitate and ease the road for future research to work on mitigation. Below is a discussion of the objectives and methodology of this research work.

As women are roses, men are the gardeners since they are the owner of the rose, either giving it life or killing it	Dehumanizing "women/roses are owned by man/gardeners "
Women are like roses, If you treat them right, they´ll bloom, if you don't, they´ll wilt.	Dehumanizing "Praising women, but this promotes stereotype "
There's a Rose Ceremony coming up and the women are handing out the roses this week.	Non-dehumanizing

Table 1. 1. Examples of dehumanizing and non-dehumanizing sentences

1.1 Research Objectives and Methodology

The overall goal of this study is to contribute to this new research area, which is the automated detection of dehumanizing language against women. We use supervised machine learning approaches to build a precise and effective tool that can systemize and

automate the detection of women dehumanizing depiction in the text. we follow the methodology displayed below throughout the study (Figure 1.1).

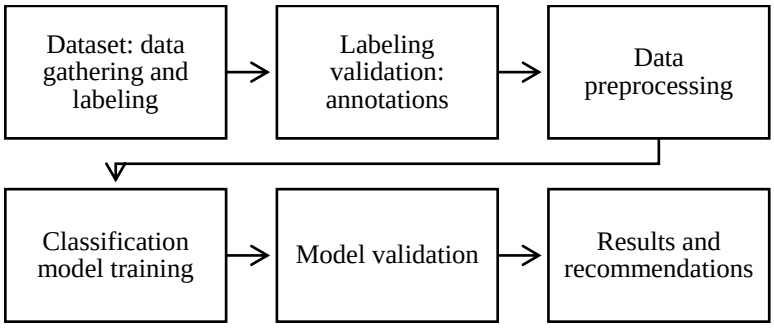

Figure 1. 1. A summary of the overall methodology for this study

The first step is building a dataset by gathering sentences from Google and Twitter. Then, we create an annotation task to label extracted sentences into dehumanizing and non-dehumanizing statements. After that, data is preprocessed using different Natural Language Processing (NLP) techniques to prepare it for the subsequent step which is training a classification model. Several classifiers are trained, tested, and evaluated towards choosing the optimal model for detecting women dehumanization in text. After that, we validate our tool by conducting a case study and implementing the developed model on a separate unseen set.

The remaining chapters are organized as follows. Chapter 2 presents an overview of the related studies done on Linguistics, Sociolinguistics, NLP, and Machine Learning. Then, in Chapter 3, a discussion on the methodology, which includes details of the process of collecting and preprocessing our data to make it suitable for model training. Furthermore, an explanation about building machine learning different models, such as logistic regression (LR), support vector machine (SVM), naïve Bayes classifiers (NB), random forest (RF), and Extreme Gradient Boosting (XGB). Chapter 4 reports

results and the best classifier after evaluating all. Finally, Chapter 5 summarizes this work, discusses its limitations, and provides pointers.

CHAPTER 2

LITERATURE REVIEW

First, for a deeper understanding of the topic of this research, a literature review is conducted to highlight what has been done and identify gaps to be filled. Therefore, this section starts with a general definition of dehumanization, metaphors, and dehumanizing metaphors. Then, it moves to an overview of the negative impact of women dehumanizing presentation on the individual level and the society.

2.1 Dehumanization Definition

Dehumanization is an arguable concept, where many studies investigate and discuss its meaning concerning numerous emphases such as ethnicity and race, gender and pornography, medicine, technology, disability, and other domains (Haslam, 2006). Nevertheless, dehumanization, in general, can be defined as denying any of the two senses of humanness: Uniquely Human characteristics (UH) and Human Nature (HN).

The UH characteristics separate humans from animals, such as civility, moral sensibility, rationality, maturity, logic, and refinement. The denial of those features or describing people by the opposite can be known as implicit or explicit Animalistic dehumanization. An example of animalistic dehumanization is proposed by Morris et al., (2018), where researchers suggest that the sexual objectification of women has some unique dehumanizing signatures. This can be demonstrated when focusing only on

females' sexual features, which strips women's humanness and represents the lack of UH attributes.

On the other hand, HN traits are central and standard to humans, such as emotional responsiveness, interpersonal warmth, cognitive openness, agency, individuality, and depth. Ignoring HN characteristics of others, like denying their curiosity, inertness, passivity, coldness, and superficiality, is a kind of Mechanistic dehumanization. Morris et al., (2018) also exemplify mechanistic dehumanization by appearance-focused objectification when emphasizing women's beauty or physical appearance, which represents the lack of HN traits.

Therefore, a person who is UH denied is seen as animal-like or subhuman, while a person who is HN denied is not seen as human, but, object-like.

2.2 Comparisons in Speech

"A metaphor is something, a simile is like something, and an analogy explains how one thing being like another help explain them both" Robert Lee Brewer

Metaphors, similes, and analogies are all types of figurative language. The three forms can be used in setting comparisons between different entities. that being the case, and for simplification, the below discussion considers all three literary devices of making comparisons in speech as metaphors.

People use metaphorical structures as implicit comparisons (Abrams, 1971). This implicitness comes from the non-literal comparison (Tourangeau & Sternberg, 1982), where the concept is considered to be metaphorical when it is used to experience and understand one thing in terms of another, and this is called conventional metaphors. Conventional metaphors are an inspiration from a human culture that is reflected in

ordinary language (Lakoff & Johnson, 1987). Researchers also introduce different kind of metaphors that is outside our conventional conceptual system. This kind of metaphor is very insightful because it can give new meanings and more creativity in understanding our activities, beliefs, and lives. It can, also, convey new perspectives and affect the recipient's thoughts (Lakoff & Johnson, 1987).

Humans tend to think metaphorically when describing a specific entity, feeling, or domain in terms of another that seems to be similar in some features (Tourangeau & Sternberg, 1982). Setting such comparisons, or bringing one term to describe something else is important for partial understanding and realizing what cannot be completely comprehended (Lakoff & Johnson, 1987). Nonetheless, the use of metaphorical phrases highlights specific features of the concept and hides others, and this suppression is not done merely or randomly but based on semantic considerations that are results or cultural views, personal ideology, and past experiences (Hines, 1999; Lakoff & Johnson, 1987). Therefore, as the selection of a metaphorical concept focuses on specific features only, it also prevents us from seeing other aspects of the concept (Hines, 1999; Lakoff & Johnson, 1987), which, consequently, might represent some semantic biases. To cite an instance, describing humans in terms of non-human items can be derogating, because this comparison, probably, highlights some features of the non-human source domain and ignores some senses of the humanness of the target domain. Thus, it is also applicable that comparing women to non-human items, to praise or slander, might represent a kind of women dehumanization. The usage of metaphorical expressions to describe women might reflect biased ideologies of the speaker, also, might convey wronged concepts and influence the conceptual system of the listener. Hence setting comparisons between

females and non-human terms could contribute to spreading dehumanization against women.

2.3 Women Dehumanizing Metaphors

Hines, (1999) and Lakoff & Johnson, (1987) explain how the use of metaphorical expressions in describing women can be derogating and belittling, specifically, when comparing females to food, animals, objects, etc. For instance, using dessert metaphors in describing females not only trivializes them to their sexuality but also equates women, subconsciously, to edible objects that can be decorated, bought, sold, and eaten (Hines, 1999). Likewise, many works in the literature demonstrate that women are belittled and seen as an object of sexuality when comparing them to animals that are usually hunted, owned, or eaten. This also puts men in a dominant position and represents women as subordinates, inferior or margin (Baider & Gesuato, 2003; Hines, 1999; López Rodríguez, 2009). A more comprehensive description can be found in (Rodríguez, 2007), where the researcher divides sexist metaphors into categories, as presented below:

1- Women as dessert:

Devaluing women and describing them as powerless and desired objects. Also, using such metaphors and describing women by adjectives that are similar to desserts adjectives, brings women derogation to the conscious level (Kang, Hye-Min; Shaydullina, 2015).

2- Women as animals including women as pets, women as farmyard animals, and women as wild animals:

Describing women by small-animal metaphors that are of lower status and are usually hunted and possessed (Kang, Hye-Min; Shaydullina, 2015), or presenting females as wild animals that must be tamed (Li, 2019).

3- Women as babies:

Babies are adults-to-be, tender, require attention all the time, defenseless, unable to do things on their own, etc. Using these metaphors to refer to females presents women as immature, needing protection, and inferiors.

4- Women as Aristocrats:

Using aristocrat metaphors to describe women places women on a level that is above ordinary men.

5- Women as supernatural creatures:

The spiritual creature used to be ranked the highest in the being chain, where those placed at the top are perfect and powerful. So, describing females as supernatural can be either praise or abuse to men.

Tarkela, (2016) develops a similar taxonomy but adjusts some sections, where the researcher starts with women as natural physical things, complex objects, and plants, where comparing women to food is considered as part of this category. The researcher, then, has the same subcategories of women as animals, aristocrats, and supernatural creatures, however, the women as babies category is part of a bigger section that is women as other people, which includes comparing women to babies and children and women as other adults.

2.4 Negative Impact of Women Dehumanizing Language

Even though our thinking affects our words, the opposite is also correct. Language plays an essential role in shaping human cognition. It influences our thoughts, memories, imagination, decisions, and even biases (Boroditsky, 2011).

One of the fundamental parts of the language system is metaphors, which are not just decorations or ornament (Erickson, 2002; Ortony, 1975), however, metaphors create a new window to making inferences, formulating knowledge about social problems and groups, and understanding complex issues (Thibodeau & Boroditsky, 2011). Moreover, metaphorical expressions can sometimes formulate the way we think of things around us. To explain, metaphorical expressions lead our thinking to focus on one specific aspect of a concept and hide other meanings (Lakoff & Johnson, 1987), which can create a new conceptual domain (Allbritton, 1995) and new meanings (Lakoff & Johnson, 1987). Semantic meanings that are driven when using metaphors in language can interfere with formulating our conceptual system that governs our thoughts and daily functioning. Thus, as our imagination is a kind of metaphorical thought, it will be affected by the partially highlighted concept that is represented when using metaphors (Lakoff & Johnson, 1987).

The selection of metaphors, to refer to or describe an object or a group of people, can reflect the speaker's cultural view and ideologies, also, metaphors can influence people's perceptions toward things and others around us (Hines, 1999; Lakoff & Johnson, 1987). This can also be applied to investigate the impact of metaphor used to describe a specific gender. Although, it is common to describe people's attitudes, appearances, life situations, etc., from both genders, using metaphorical expressions. However, females representation is usually associated with trivialization and negative

connotation, more than men (Baider & Gesuato, 2003; Hines, 1999; López Rodríguez, 2009; Rodríguez, 2007).

Starting with the hunger metaphor, it is common that human desire (desire for love, eating, power, sex, …etc.) is linked to hunger, where there is a known metaphor DESIRE IS HUNGER (Lakoff & Johnson, 1987). It is, also, common to use terms of eating and feeling hungry to express human sexual desire and sexual satisfaction (Baider & Gesuato, 2003).

Therefore, using dessert and food metaphors to describe women, equating females to edible objects, might influence the perception of womanhood. For instance, describing females by dessert metaphors (e.g. cheesecake, pie, tart, and other food that is not essential to our diets) brings the idea of the unimportance of women (Hines, 1999). Also, dessert is usually delicious, mouthwatering, eaten, sold, owned, and tends to be decorated (Hines, 1999), all of these features highlight the idea that women need to be visually appetizing and use make-up, for example, to improve outward appearance (Eble, 1996). All in all, the metaphorical representation of women might bring women devaluation and derogation to the conscious level (Kang, Hye-Min; Shaydullina, 2015) and affect our perception of females and their roles in society.

Moving to animalistic metaphors, depicting people as animals or objects, in general, conveys negative connotations and human degradation. This is usually explained by the hierarchical organization of the Great Chain of Being (López Rodríguez, 2009).

Thus, animalistic metaphors are a kind of derogation tool to degrade a particular social group and to highlight only undesirable characteristics to set this group as inferior. Women, in general, are more likely to be animalized (Goldenberg et al., 2009;

Heflick & Goldenberg, 2014), this might be occurring because of the physiological nature of females " fertility, breastfeeding, reproductivity, etc." (Reynolds & Haslam, 2011; Rudman & Mescher, 2012). Whereas, even when both genders are compared to animals, still men are described as in the dominant position and women as subordinates, with a more negative connotation when referring to women. In addition, animalizing women is linked to sexual objectification and lacking UH attributes (Morris et al., 2018; Morris & Goldenberg, 2015), where referring animalistic metaphors to women represent them as an object of sexual desire and describe them as subordinate animals with minimal power (Baider & Gesuato, 2003; Hines, 1999). To explain, using pet metaphors show females as sweet, kind, weak, and cute, which, implicitly, means that they have to be kept under the dominance of men (Li, 2019). On the other hand, describing women by wild animal metaphors is usually associated with traits such as fearless and that woman is subject to taming, which keeps the ideology of the dominant man (Li, 2019).

The worse impact of using these metaphors is that they can affect human actions and future functioning. In particular, linking a female to animals, representing her lack of HN attributes, and associating this with the idea that females feel pain less (Morris et al., 2018) might be a reason for violence against women. Tipler & Ruscher, (2017) examine the ability of animalistic metaphors in shaping sexist attitudes against women. In their experiment, researchers ask participants (males and females) to complete two seemingly unrelated studies. The first is a fake political report about the importance of female votes in the election seasons. The second is the main study, which aims to examine appearance-based and behavior-based impressions. Both studies are formulated in three versions (using women-as-prey metaphors, women-as-predator metaphors, and

human baseline). Participants are also asked to complete some distractor questionnaires before introducing the Ambivalent Sexism Inventory (ASI). The ASI test (Glick & Fiske, 1996) includes 22 statements that assess the acceptance of culturally transmitted against women attitudes and reflect hostile and benevolent types of sexism. Results demonstrate that using "women-as-predator" and "women-as-prey" metaphors can shape hostile and anti-female sexist attitudes. In addition, once metaphors that are relevant to gendered power are presented, conflicts come to appear, perpetuating the harmful beliefs and stereotypes about women's roles in society (Tipler & Ruscher, 2017).

Besides, Rudman & Mescher, (2012) investigate the association between dehumanizing women as animals and objects and the sexual aggression of men. In their first study, they use the Implicit Association Test (IAT) to investigate the association between humans and animals, where participants are asked to categorize words (women, woman, female, she, her, girl and men, man, male, him, he, boy) with either animal characteristics (animals, nature, instinct, physical, bodies) or human characteristics (culture, society, mind, symbols, monuments). Researchers also use the Ambivalent Sexism Inventory (ASI) to assess hostile sexism and benevolent sexism toward women, by rating the ASI items from 1 (strongly disagree) to 5 (strongly agree). Additionally, sexual harassment is assessed using the Likelihood to Sexually Harass (LSH) Scale, which comprises 10 vignettes to make participants imagine that they have power over each other, and they can use this power to coerce another participant into having sex, scaling from 1 (not at all likely) to 5 (very likely). Furthermore, a subset of items from the Attraction to Sexual Aggression (ASA) Inventory (Malamuth, 1989) is used to measure interest in consensual sex and rape proclivity, prompting in each item

"If you could be assured that no one would know and that you could in no way be punished for engaging in the following act, how likely, if at all, would you be to commit the such act?". Finally, Rudman & Mescher, (2012) use the Attitudes Toward Rape Victims Scale (ARVS; Ward, 1988) to assess the negative attitudes, such as trivialization, victim blaming, and deservingness. While, in the second study, Rudman & Mescher, (2012) add two men and women objectification types, comparing them to objects and animals. As well, a rape-behavioral analog (RBA) (Widman & Olson, 2011) to measure male sexual aggression more directly, by obliging men to decide between violent and sexually violent images to present to females, purportedly, for a project. Results of the two studies show that dehumanizing women is automatically related to men's sexual aggression. Where objectifying women involves treating them as a tool for men's purposes, especially, with the denial of women's feelings, making females more likely to be violated and victims to rape (Rudman & Mescher, 2012)

Similarly, Bock & Burkley, (2019) experiment on the impact of "men-as-predator" and "women-as-prey" metaphors on rape proclivity and sexual violence against women. Researchers ask participants to read a vignette that either depicts a man pursuing a woman or the opposite and in either metaphor condition (including predator/prey metaphorical expressions) or the control condition (identical vignette but with non-metaphorical phrases). After reading the vignettes, participants are asked to complete Illinois Rape Myth Acceptance Scale (IRMA) items (IRMA; Payne et al. 1999; McMahon and Farmer 2011) to assess attitudes toward rape perpetrators and their victims. Participants, also, complete 5 items from the Attraction to Sexual Aggression scale (ASA). Moreover, (Bock & Burkley, 2019) apply another experiment, where they use Amazon Mechanical Turk (Mturk) to collect data and include IRMA and ASA

measures with the same vignettes, excluding the women-pursuing-a-man vignettes. Furthermore, a new measure is added that is making participants read about rape scenarios and then indicate how likely they would act the same way the perpetrator acts in the situation, this is to assess potential reasons for the rape proclivity. Results show that the metaphorical depiction of men-as-predator and women-as-prey in the romantic context (specifically heterosexual relationships) affects men's attitudes toward rape by increasing their rape myth acceptance and rape proclivity (Bock & Burkley, 2019).

Collectively, the above discussions show how we – humans – are our words. Whether speaking, writing, reading, or listening to any form of language, it might have significant impacts on the way we feel toward things and people around us, and the way we think and believe in specific thoughts, stereotypes, or ideologies. The language might even affect our actions and readiness to act in harmful or harmless ways.

2.5 Extracting Women Dehumanizing Metaphorical Expressions

A series of existing studies investigate dehumanization and dehumanizing metaphors with gender, in general, and to females, specifically. Most of these works used manual approaches for detecting metaphorical phrases to be later analyzed, which represents the detection stage as a key for further stages of the research. The manual approach is reading full texts and analyzing them on the word level. For instance, researchers extract full articles from popular women's and teenage magazines and websites, such as *CosmoGirl*[1], *Vanity Fair*[2], *Top of the Pop*[3], *Seventeen*[4], and

[1] https://en.wikipedia.org/wiki/CosmoGirl
[2] https://www.vanityfair.com/
[3] https://en.wikipedia.org/wiki/Top_of_the_Pops_(magazine)
[4] https://www.seventeen.com/

cosmopolitan[5]. After that, qualitative approaches are followed to analyze the basic and contextual meanings of metaphors (Rodríguez, 2007; Tarkela, 2016). Rodríguez, (2007) also selected articles from English and Spanish written press to analyze women as dessert metaphors. Where metaphorical phrases are manually extracted from several magazines, such as *People*[6], *Vanidades*[7], *La Vanguardia*[8], *Rogazza, USA Today*[9], *New Yorker*[10], and others. Recently, López Maestre, (2020) in their investigation on women and hunt metaphors, Google search engine is used for material gathering, yet, it is still manual. To explain, after deciding on a lexical list of terms that can be used metaphorically, Google search is obtained to look for different texts, such as books and articles, through requests like "books on husband hunting" (López Maestre, 2020), and qualitative analysis then proceeds. Similarly, (Vujković & Vuković-Stamatović, 2021) use Google search to obtain women's physical appearance and animal metaphors from Montenegrin webpages.

2.6 Machine Learning

Very few works in literature are done on the field of metaphor detection, in general. Schulder & Hovy, (2015) proposes an approach to detect metaphors through term relevance to measure how a word is out of place or unusual in a given context. Measuring the unusualness of words is done via domain-specific term relevance metric, which consists of domain relevance and common relevance. To explain, domain relevance is based on *term frequency inverse document frequency* (TF-IDF), which is

[5] https://www.cosmopolitan.com/
[6] https://people.com/
[7] https://en.wikipedia.org/wiki/Vanidades
[8] https://www.lavanguardia.com/
[9] https://www.usatoday.com/
[10] https://www.newyorker.com/

used to measure the impact of each term on a given text, so terms with a low score in this feature can be considered metaphors. However, as TF-IDF gives low scores to very frequent terms among all domains, normalized document frequency is used as a common relevance indicator to filter out common terms. After classifying all the words as binary labeling, machine learning classifiers are used, which are the Hidden Markov model (HMM), Conditional Random Fields (CRF), and Support Vector Machine (SVM).

Prior research done by Recasens et al., (2013) introduce a linguistically-informed model to automatically detect biased language. To identify biased words, the *neutral point of view* (NPOV) policy, which is advocated by Wikipedia, is used. Thus, based on NPOV policy, a list of words that must be avoided to ensure fairness is prepared to train the model. Where the researcher trains a logistic regression model to identify biased words according to their probabilities (Recasens et al., 2013). Moreover, some studies also focus on topics that are close to dehumanization and its consequences. ElSherief et al. (2018) propose methods to detect hate speech, anti-social behavior, online harassment, and cyberbullying using Sparse Additive Generative Models (SAGM) for topic modeling, along with sentiment analysis. In addition to toxicity and social bias detection using hybrid classification and language generation tasks (Sap et al., 2019). The first computational framework for analyzing dehumanizing language is introduced by Mendelsohn et al., (2020). The proposed framework is applied to *New York Times* articles over a period of time to analyze dehumanizing language towards LGBTQ people. Mendelsohn et al. (2020) use word2vec models to figure out how LGBTQ terms are represented semantically within these models, along with proposing

some quantitative approaches. Nevertheless, there have been no proposed approaches or tools to automatically detect the dehumanizing depiction of women in the text.

To summarize, a closer look at the literature on women dehumanizing presentation in English reveals that the focus is on the linguistic and social aspects of the topic. To be specific, existing research works mainly focus on understanding the comparative and metaphorical language, manually extracting metaphors, categorizing them, analyzing the semantic meanings, and figuring out the negative impact on society. Considering the complexity of analyzing gender bias in semantics, an arguably important question to be addressed is how to accurately define and extract dehumanization from the text? We therefore will focus on building a detection system to fill this gap and pave the way for future research. Toward this goal, we will follow the methodology described in the next Chapter.

CHAPTER 3

METHODOLOGY

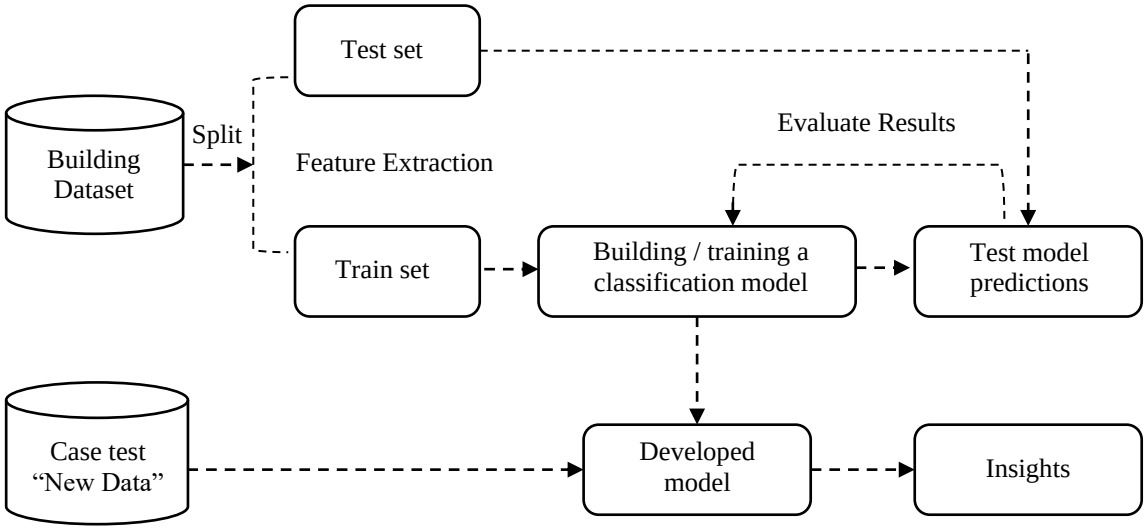

Figure 3. 1. Summary of our methodology

This Chapter presents the process of our experiment, starting with data preparation, going through building machine learning models and ending with validating our final model (Figure 3.1). The software used to implement our analysis and build the proposed tool is python programing language, where libraries used are:

- NLTK and RE: for text cleaning and normalization.
- Sklearn: for text preprocessing and modeling.
- Grid Search CV: for model hyperparameters tuning.
- Numpy, Pandas, and Matplotlib: for results visualization and evaluation.

3.1 Data Preparation

According to our knowledge, there is no existing dataset on sexist metaphors and figurative language that might be usable in training an accurate machine learning model to automate neither the detection of the dehumanizing depiction of women in text nor mitigating it. Therefore, a key contribution of this study is developing a labeled dataset that can serve the purpose of our study and will also be useful for future research, through the below steps (Figure 3.2).

3.1.1 Data Collection

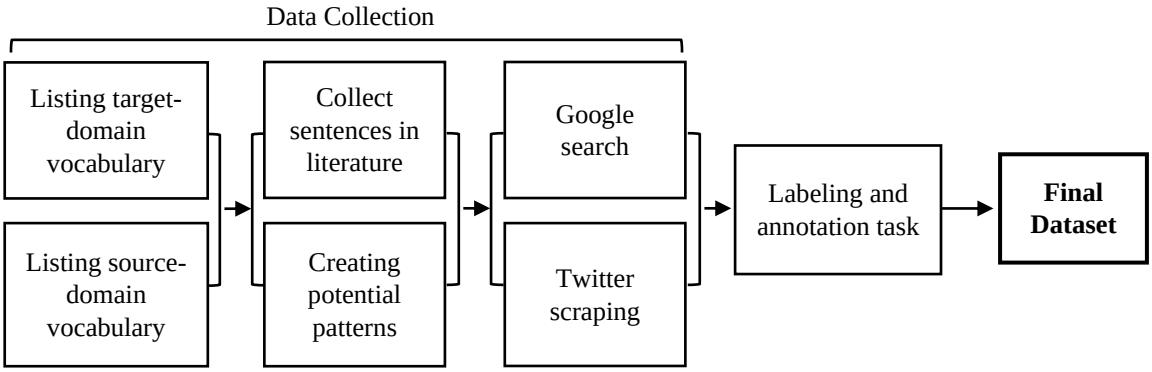

Figure 3. 2. Process of building the dataset

3.1.1.1 Metaphors Taxonomy

In our research, we adopt one of Stefanowitsch, (2008)'s suggested strategies. These strategies help in searching for metaphorical expressions for automatic and semi-automatic detection. The proposed approach is to, first, select all target-domain vocabulary, search for the source-domain vocabulary, and list all potential terms. After that, the searching process is done based on patterns that are formed by the collected terms (Stefanowitsch, 2008). Accordingly, as the target in this study is females, thus,

the selected target-domain terms are (she, her, girl, girls, lady, ladies, woman, and women). Our next step is to identify the source-domain terms. For this purpose, we combine the introduced taxonomies by both Rodríguez, (2007) and Tarkela, (2016). Moreover, we use the metaphors lists that are defined by Hines, (1999), (1999) for describing women as desserts and animals. Additionally, metaphors that (Aslan, 2015, 2021) lists as a result of their experiments are added. Our final taxonomy and selected source-domain vocabulary are presented below (Table 3.1).

The next step is to collect sentences that are mentioned in previous studies. Rodríguez, (2007), (2007) analyzes a set of metaphorical expressions about women that are mentioned in teenage and women's magazines. Besides, Aslan, (2015), (2021) sets two experiments asking teachers to describe men and women using figurative structure using the pattern " women are …because … / men are …because …" . All the mentioned statements from all these research works are collected and added to our dataset.

3.1.1.2 Patterns Creation

As suggested by Stefanowitsch, (2008), to find our targeted type of sentences, we need to create patterns that could include terms from either source-domain only or both source-domain and target-domain. Therefore, through analyzing the collected sentences, some sentences are found to have consistent patterns with the Stefanowitsch, (2008)'s strategy. For instance, Aslan,(2015), (2021) in their experiments use the template of "women/men are * (a metaphor) because …" and ask participant to fill the gap by a metaphor of their choice. The pattern of those sentences is "women are *

metaphor". For metaphorical sentences that have no clear pattern, patterns are created using the selected lexical items from the target domain and/or source domain.

Metaphor Category	Source-domain Terms	Example
Women as food	Cookies – tough cookies – sweet chocolate – jam – pie – cutie pie – crumpet – apples – salt – rib – banana – cheesecake – brown sugar – strawberries – tart – cherry – honey – pumpkin – sugar – plum	*was she hot? Nah man she was brown sugar, good all over*
Women as plants	Roses – flowers – tree	*Women are like roses whose beautiful flowers when they are just opened, are wrinkling*
Women as objects	Glass – wine glass – tea bag – snowflake – snowdrop – candles – water – sun – soil – river – moon – ass – hoes - computers – vase – books – machines – vessel	*A woman is like a tea bag, you can't tell how strong she is until you put her in hot water*
Women as insects	Bees – butterflies – ants	*Women are like bees, sweet and poisonous*
Women as animals	Swan – parrot – bitch – vixen – fox – foxy – chicken – chick – cat – catty – pussy – kitty – bird – goat – hen – game hen – kitten – lioness – rabbit – mare – cow – lamb – sex kitten – canary – snake – prey – animal	*How could he marry a snake like this*
Women as other people	Baby – child	*Work it, baby! Make your Saturday job work for you*
Women as aristocrats	Queen – princess	*She is a queen of the mattress*
Women as supernatural creatures	Angel – vamp – goddess – sex goddess	*Embrace your inner sex goddess and the men will fall at your feet.*

Table 3. 1. Women dehumanizing metaphors taxonomy

To explain, we mainly use two methods for stating the patterns:

1- Patterns that have vocabulary from both target-domain and source-domain (Stefanowitsch, 2008): we select a specific metaphor “e.g. butterfly” and start to iterate the target domain items for each search round (e.g. “she is * butterfly”, “women are * butterflies”, “girls are * butterflies”, “she * like butterfly”, “women * like butterflies”, etc.).
2- Some metaphors are usually used when referring to women, however, the sentence does not state women or females explicitly (using a term from target-domain vocabulary). Thus, for those metaphors, we try to use only the source domain to create the pattern with an accompaniment adjective or a word (e.g. “ * tough cookies”, “ * sexy chicks”, “ * sex goddess”, “ * cutie pie”, etc.)

3.1.1.3 Building our Dataset

All the prior steps, like listing lexical units from the source domain and target domain, also, identifying potential patterns of the sentences, are done to ease our next step, which is collecting targeted sentences. Data are collected from two different sources:

1- Google search: for each search round we use one of the created patterns as a search criterion, then, we collect all the resulting sentences, whether they are dehumanizing or not. Sentences that are collected using Google search are around 1700 sentences. These statements are manually labeled to facilitate collecting more data in the second search round using Twitter API.
2- Twitter scrapping: similarly, via Twitter API using the “tweepy” python library, we scrap tweets using our patterns as a search query. The total number of

extracted tweets is around 1400 tweets. Scrapped tweets are to be manually labeled by annotators.

After removing all duplicates, our final dataset contains a total of 3011 labeled sentences.

3.1.2 Labeling Task

The collected data consists of dehumanizing and non-dehumanizing phrases. Where all sentences include lexical items from the metaphors and have similar patterns. Some of these sentences are very clear to be either dehumanizing or non-dehumanizing. Whereas a considerable number of statements are less clear-cut and semantically dehumanizing, which makes the labeling process arguable, and labels might differ from one person to another. For that reason, as a means to make our tool neither too sensitive nor ineffective and as human-like as possible, the manual annotation of our data is done by participating annotators.

To ensure that the labeling task is valid, we have selected participating labelers who have an excellent English level as they are all English curriculum educated. They, also, either have an interest in gender issues or have studied gender studies courses. Annotators (including us) are 11, where 8 are females and 3 are males.

3.1.2.1 Labelling Protocol

Annotators are, each separately, given the whole dataset for the labeling task. Where the classification is binary (1 | 0), and variables are defined below (Figure 3.3).

- Non-dehumanizing statements 0: non-dehumanizing sentences are statements that include one or more of the listed source-domain lexical items but present only the original and literal meaning of the word and do not deny any of UH or HN attributes.
- Dehumanizing statements 1: dehumanizing sentences are statements that include one or more of the listed source-domain lexical items but present a semantic meaning of the word and does, implicitly or explicitly, deny any of the UH or HN attributes

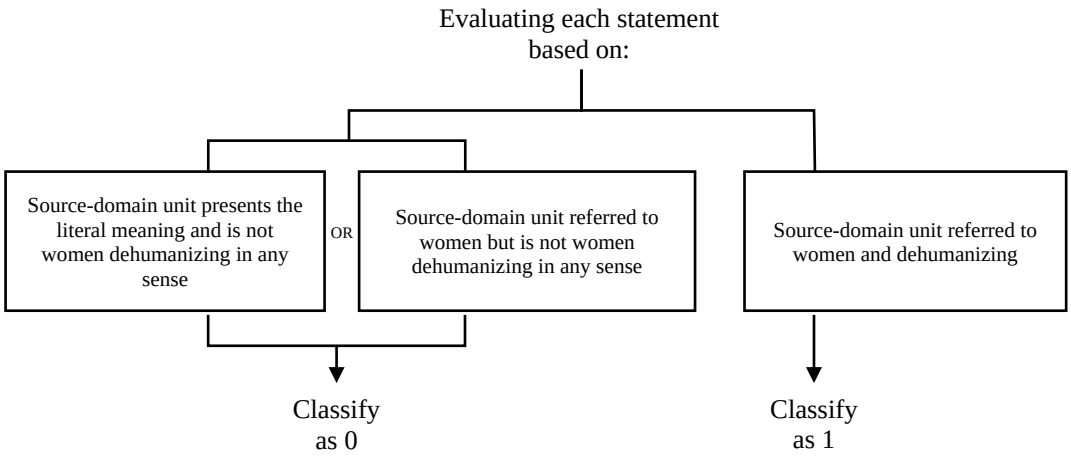

Figure 3. 3. Labeling process summary

After each annotator is done with labeling collected sentences, deciding on the final label is our next step. The final label is chosen based on the majority of votes. In other words, there are 11 votes for each statement, if 6 or more votes are agreed to be 1, the sentence is dehumanizing and finally labeled as 1. Alternately, if 6 or more votes agreed to be 0, then the sentence is non-dehumanizing and finally labeled as 0. On this basis, our final dataset contains a total of 3011 sentences, where 1528 of them are dehumanizing and 1483 are non-dehumanizing. (Figure 3.4)

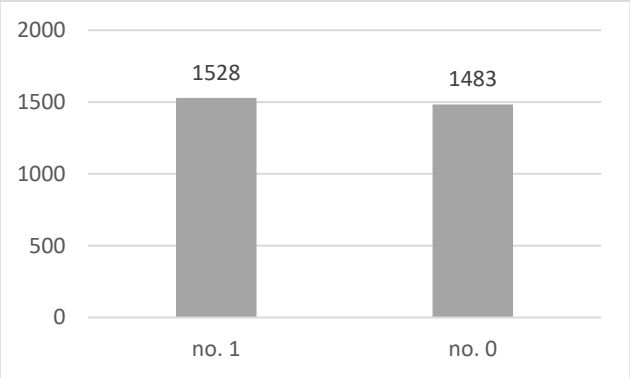

Figure 3. 4. Visualization of our dataset

3.1.3 Data Splitting

As displayed in Figure 3.4, our dataset is almost balanced, Thus, before building and training classification models, data is split into 80% (2408 records) for training the models and 20% (603 records) for the testing step. This splitting ratios are fixed for training/testing of all the implemented classifiers.

3.1.4 Data Pre-processing

Collecting good data and, properly, cleaning and preprocessing it is as important as building machine learning models. That is because training a model on bad data will result in undesirable performance. Besides, when using real-world data, this data may be noisy, irrelevant, redundant, incomplete, etc. (Kuhn & Johnson, 2013). Which highlights the importance of doing data preprocessing before starting the modeling stage (Quilumba et al., 2014). Nevertheless, our data is not randomly collected, and it is less likely to be irrelevant, however, there is a possibility that cleaning and normalizing the text might give better results. As described next, several combinations of NLP techniques are implemented to reach the best for our models.

3.1.4.1 Duplicates removal

Duplicated records in datasets are problematic and can affect the efficiency of machine learning models for multiple reasons. First, violating the independence of the training data that is caused when having identical entries in the training and testing sets. Which might cause a misleading or biased performance of machine learning models. Additionally, identical records might lead to higher performance, while in reality, this is not correct because models are not dealing with new but frequent data. Therefore, removing duplicates is done before attempting any NLP preprocessing technique.

In our case, after sentences are collected from literature, Google, and Twitter, the first built dataset contains around 3100 records. By removing the duplicates, we are left with 3011 sentences.

3.1.4.2 Expanding contractions (CE)

A contraction is a short form of a word or group of words. When two or more words are put together, they can be shortened and form a new word, a contraction (Webster, 2021). For instance, **isn't** (one token) stands for **is not** (two tokens) and **don't** (one token) stands for **do not** (two tokens). Machines understand contractions as new words and different from the original ones, which is why expanding contractions might result in better performance (Hapke et al., 2019).

3.1.4.3 Lowercasing (LC)

Machines interpret words like cat and Cat differently. Thus, when all data is in the same case it makes it less complex for machines to treat (Egger, 2021). Usually, in machine learning, lower case is preferred, however, we need to try all options and evaluate models' performance, especially, when uppercased words have different meanings. For example, we have the metaphor "rose", which can be used as the name "Rose". Consequently, uppercasing the word will affect the meaning of the sentence and result in a different classification.

3.1.4.4 Punctuation removal (PR)

Common punctuations are around 32, which are "!"#$%&'()*+,-./:;⇔?@[\]^_'{\}~". These marks are usually used for purposes like, e.g. dividing the sentences or giving some sentiments to the sentence. Moreover, machines treat the word **hello** and **hello!** differently. Therefore, removing punctuation makes each text treated equally (Egger, 2021).

3.1.4.5 Stop words removal (SW)

Stop words are a collection of commonly used words that are highly repetitive in speech, such as "a, an, the, so, yet, but, so, etc.". Usually, these words add no valuable information to the analysis because they are highly repetitive and are not considered keywords (Porter, 2006). However, in our case, we need to try all options as adding (e.g. not) in the sentence might change its label. For example, the sentence "women are snakes" is dehumanizing, while "women are not snakes" could be non-dehumanizing.

Therefore, including stop words might play an important role in clarifying the sentiments of the sentence.

3.1.4.6 Lemmatization (L) and Stemming (S)

Lemmatization and stemming are two similar techniques that aim to break a word down to its root (Méndez et al., 2006). Lemmatization takes into consideration the word's part of speech, then replaces it with its lemma. For instance, lemmatizing the word "caring" will be "care". On the other hand, stemming a word reduces a term to the root without considering the part of speech, which might lead to a new meaning. For instance, stemming the verb "caring" will be "car" (Egger, 2021). In our analysis, we try both techniques separately to figure out the best.

For our analysis, we create several preprocessing pipelines of the same data, each version is differently preprocessed. Iterating the model over each of the pipelines will help us know which preprocessing technique gives higher performance. The pipelines are:

1- Raw: raw data.
2- CE/PR: contractions expansion and punctuation removal.
3- CE/PR/LC/SW: contractions expansion, punctuation removal, lowercasing, and stop words removal.
4- CE/PR/LC/SW/S: contractions expansion, punctuation removal, lowercasing, stop words removal, and stemming.
5- CE/PR/LC/SW/L: contractions expansion, punctuation removal, lowercasing, stop words removal, and lemmatization.

3.1.5 Feature Extraction

Raw textual data is very unstructured and complex for machines to deal with because machines do not understand words but numbers. Therefore, to feed texts to machine learning algorithms, texts must be converted into numerical form (a matrix or vector) of features via feature extraction techniques.

In our study, we use two common feature extraction methods that are usually used for text classification:

- Bag of words (BoW).
- Term frequency-inverse document frequency (TF-IDF).

3.1.5.1 Bag of Words (BoW)

The BoW model is a popular method that is usually used for vectorizing text documents (Harris, 1954). The core of this method is to express sentences through a matrix of features. To explain, each unique word is assigned a separate column, and each entry in the matrix marks the presence of the word 1, or its absence 0 (Table 3.2). This way, there will be a set of unordered frequencies for each unique word in a document (Salton, G., Buckley, 1986).

3.1.5.2 Term Frequency-Inverse Document Frequency (TF-IDF).

The TF-IDF approach is best known for weighing words according to their uniqueness. In other words, this method captures the relevancy of document words to particular categories (Y. T. Zhang et al., 2005). To explain this approach in more detail, the TF-IDF will be broken down as shown below:

Term Frequency (TF): this calculates the proportional frequency of a term relative to the entire document, which is formulated as:

$$tf(w,d) = \frac{no.of\ times\ w\ occurs\ in\ d}{total\ no.of\ words\ in\ d}$$

Inverse Document Frequency (IDF): this measures how rare or frequent a term is across all documents of the entire corpus, where rare words have a high IDF score.IDF is calculated by :

$$idf(d,D) = log\ (\frac{total\ no.of\ documents}{no.of\ documents\ containing\ the\ word})$$

Term Frequency-Inverse Document Frequency (TF-IDF): is calculated by the multiplication of TF by IDF, and formulated as:

$$Tfidf(w,d,D) = tf(w,d) * idf(d,D)$$

In our study, we implement TF-IDF on the word level, besides, using n-gram (1,2) to take the range unigram (only single word), and bigram (group of two words in a row) to figure out the best performance.

3.2 Modeling

3.2.1 Classification Models

The current study introduces a work that has not been done yet, which is to automate the classification of whether a sentence has a dehumanizing presentation of women or not in an English text. For this purpose, as building a comprehensive dataset to train a model is an essential step for building accurate classification models, we have built a dataset that contains over 3000 sentences and over 60 metaphors. Moreover, all

sentences have been labeled by 11 annotators to ensure neutrality and efficiency of our work.

Our next phase is training different classification models, testing their performance, and evaluating their performance on new real data to judge whether the deployed model is effective and accurate in classifying new texts or not.

In this study, five classifiers are implemented and evaluated, logistic regression, naïve Bayes, support vector machine, random forest, and extreme gradient boosting. The start is building a logistic regression as a base line model, where it is iterated over the five preprocessing pipelines to evaluate its performance with each of them. Then, the five classifiers run on the pipeline that reflects the best performance of the baseline model. All performances are evaluated and compared to each other to choose the most accurate performing model. Furthermore, a case study is designed to evaluate the performance of the best model on new unseen data.

3.2.1.1 Logistic Regression (LR)

Logistic regression is one of the most popular methods that are used for the analysis of binary events, in other words, when having a binary response (Hilbe, 2009; Jill C., 2011). Where this method is used in all the social sciences (Hilbe, 2009). The formulation of the LR model determines the strength of any feature through a process that is known as estimating the maximum likelihood (Hosmer & Lemeshow, 2013), as the model calculates the binary occurrence probability of an event (yes / no). Furthermore, the usage of the LR algorithm in machine learning can be considered an example of supervised learning, which makes this method applicable to our study.

3.2.1.2 Multinomial Naïve Bayes (MNB)

Naïve Bayes is a probabilistic machine learning model that is widely used for clustering and classification (Lowd & Domingos, 2005). Naïve Bayes classifier assumes conditional independence of the variables or features, and that contributions of all features are equal (Eyheramendy et al., 2013), which means that the presence of one predictor does not affect the others. In our experiment, we use the Multinomial Naïve Bayes (MNB) as it is mostly used for text classification, where the model takes the frequency of words appearing in a document as the feature or predictors.

3.2.1.3 Support Vector Machines (SVM)

Support vector machine is a method that has been significantly successful in learning tasks, where this method is used for classification problems and can result considerably in high classification accuracy (Tong & Koller, 2001). The main idea of the SVM algorithm is to find the optimal hyperplane in N-dimensional space (N is the number of features) to best classify the data items (Figure 3.5). Thus, to minimize the risk of wrong classifications of future data points, the plane should have the maximum possible margin (i.e the maximum distance separating observations of each class) (Y. Zhang, 2012).

3.2.1.4 Random Forest (RF)

Random forest classifier consists of a set of ensembles, which are individual tree classifiers. Where the tree classifier is a flowchart that is structured as a tree, internal nodes of the tree represent a test on a specific attribute, branches represent the outcome of the test, and terminal nodes hold the class label. Each tree in the RF votes for a class

prediction and the final prediction of the RF model is determined by the majority voting of each tree predictor. This method can achieve significant improvements in accuracy and will outperform any individual tree classifier (Breiman, 2001).

3.2.1.5 Extreme Gradient Boosting (XGB)

The Extreme Gradient Boosting model is an implementation of a technique that is known as the Gradient Boosted Trees. The idea of Gradient Boosting is that each tree predictor gets trained on the predecessor residual error. For further clarification, in the XGB algorithm, a consecutive form of a decision tree is created and then fed all the explanatory variables. The weight of variables that are wrongly predicted is increased and then fed to the next decision tree. The ensemble of all individual classifier trees gives the preciseness of the model (Chen & Guestrin, 2016).

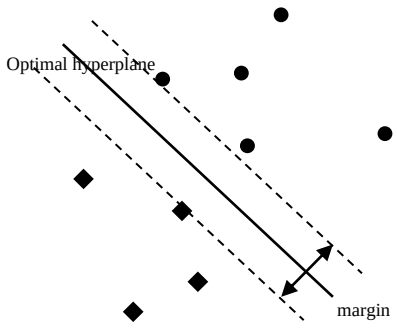

Figure 3. 5. The optimal separation hyperplane of the SVM classifier

3.2.2 Model Tuning

Machine learning models have several Hyperparameters which are usually chosen based on the previous trials before training the model. Whereas different data samples might not have the same best hyperparameters. That being the case, we use Grid search CV, which is a searching method that iterates over specified parameters and evaluates different parameter combinations to find the best hyperparameters based on our features.

3.2.3 Evaluation of Classification Models

For the evaluation of the selected classification models' performance, several measurements are implemented throughout our analysis.

3.2.3.1 Confusion Matrix:

The confusion matrix is a key performance measure for machine learning classification tasks. It represents four combinations of truly predicted and falsely predicted values. As illustrated in Table 3.3, the confusion matrix displays the four values:

- True Positive (TP): it is the number of dehumanizing sentences that are correctly classified by the model as dehumanizing.
- True Negative (TN): it is the number of non-dehumanizing sentences that are correctly classified by the model as non-dehumanizing.
- False Positive (FP): it is the number of non-dehumanizing sentences that are incorrectly classified by the model as dehumanizing.

- False Negative (FN): it is the number of dehumanizing sentences that are incorrectly classified by the model as non-dehumanizing.

	Actual Values		
Predicted Values		Positive (1)	Negative (0)
	Positive (1)	TP	FP
	Negative (0)	FN	TN

Table 3. 2. An illustration of the Confusion Matrix

The importance of this measure is that it is essential for calculating the other criteria, such as ROC/AUC curves, accuracy, recall, precision, and specificity.

3.2.3.2 ROC curve (Receiver Operating Characteristic curve):

This measurement can summarize the overall performance of the model as it includes many other criteria into it. To explain more in detail, the ROC curve is a graph that shows the True Positive Rate and False Positive Rate of a model's classifications at all classification thresholds, where the higher the curve the better the classifier's performance (Figure 3.6). The represented parameters are defined as follows:

- True Positive Rate (TPR) (i.e., Recall / Sensitivity): which is also known as recall and it represents the ratio of sentences that are correctly predicted as dehumanizing out of the total number sentences that are actually dehumanizing. It is calculated by dividing the total number of true positive classifications over the total number of positive observations, and formulated as:

$$TPR = \frac{TP}{TP+FN}$$

- False Positive Rate (FPR): this represents the ratio of sentences that are incorrectly classified as dehumanizing out of the total number of sentences that are actually non-dehumanizing. It is formulated as:

$$FPR = \frac{FP}{FP+TN}$$

In this study, the aim is to get a higher TPR with less FPR, which can prove that the model can detect dehumanizing sentences correctly.

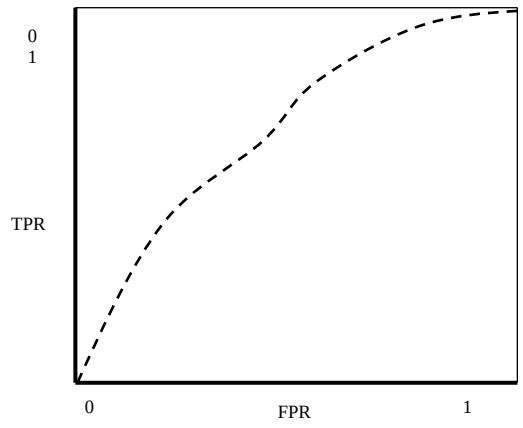

Figure 3. 6. Visualization of ROC curve

3.2.3.3 Area Under Curve (AUC) score:

AUC is the shaded area under the ROC curve in Figure 3.7. That is, the AUC score represents the entire area beneath the ROC curve, and it measures how good the

model is in distinguishing between different classes, or in other words, the degree of separation.

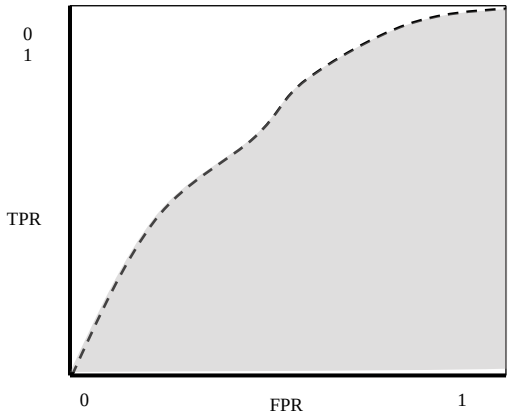

Figure 3. 7. Visualization of AUC / ROC curve

3.2.3.4 Accuracy:

Accuracy defines how correct the model is in classifying the observations. To rephrase it, it measures the percentage of correct classifications out of the total predictions and is formulated as:

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN}$$

3.2.3.5 Precision:

The precision measure indicates the quality of the model's positive predictions, which means that an effective model with high precision can precisely detect dehumanizing phrases. It is calculated by:

$$Precision = \frac{TP}{TP+FP}$$

3.3 Test Case

To test our model's behavior and the learned logic, we introduce a case study that contains raw texts that have not been seen by the model before. Testing the model on a new dataset not only helps in evaluating the functionality of the developed tool but also, helps in investigating and identifying the critical errors that could be occurring in any of the methodology stages.

The case study is built by collecting the lyrics of rap songs. The selection of rap songs is based on articles published in American women's magazines and music websites. Moreover, to make sure that these songs include some women dehumanizing sentences, we select articles about misogynistic and derogating rap songs, in addition to songs that may not have the same aggressive language, yet, about women (Table 3.4). The articles are:

1- "12 Songs with Lyrics That Are Actually Super Misogynistic*"* from *Bustle*.[11]

2- "The Most Disrespectful. Misogynistic, Dehumanizing, Inglorious Rap Songs that Black Women Love!" from *WordPress.com.*[12]

3- "Degradation of Women in Hip-Hop Music Lyrics" from *Google Sites.*[13]

4- "25 Rap Songs About Women" from *The BOOMBOX*.[14]

[11] https://www.bustle.com/articles/137558-12-songs-that-are-actually-full-of-super-misogynistic-lyrics
[12] https://empoweringyoungbrothas.wordpress.com/2013/01/10/the-most-disrespectful-misogynistic-dehumanizing-inglorious-rap-songs-that-black-women-love/
[13] https://sites.google.com/site/hiphopmusiclyrics/top-10-degrading-songs-towards-women
[14] https://theboombox.com/25-rap-songs-about-women/

	Song Name	***Singer***	***Source***
1	Blurred lines	Robin Thicke and Pharrell Williams	**Bustle**
2	So much better	Eminem	
3	Gold Digger	Kanye West	
4	Better than revenge	Taylor Swift	
5	Fine China	Chris Brown	
6	Ain't no fun	Snoop Dogg	
7	U.O.E.N.O	Rocko Featuring Rick Ross and Future	
8	It's so easy	Guns N' Roses	
9	Talk dirty	Jason Derulo	
10	Bitches ain't shit	Snoop Dogg	
11	Love game	Eminem and Kendrick Lamar	
12	One less bitch	N.W.A.	
13	I'm a dog	Gucci Mane	**WordPress**
14	Wait (The whisper song)	Ying Yang Twins	
15	Get low	Lil Jon and The East Side Boyz	
16	Blow the whistle	Too Short	
17	Choosin'	Too Short	
18	Shake that monkey	Too Short	
19	Hoochie mama	2 Live Crew	
20	Becky	Plies	
21	Tipdrill	Nelly	
22	Give me that	Webbie	
23	I get around	Tupac Shakur	
24	Hootie hoo	OutKast	**Google Sites**
25	Smack that	Akon	
26	Shake that	Eminem and Nate Dogg	
27	No hands	Roscoe Dash and Waka Flocka Flame	
28	Alphabet bitches	Lil Wayne	
29	I wanna fuck you	Akon	
30	Heidi hoe	Common	
31	Wildflower	Ghostface Killah	
32	Bitties in the BK Lounge	De La Soul	**The BOOMBOX**
33	Around the Way Girl	LL Cool J	
34	Girls, Girls, Girls	Jay-Z	
35	Ms. Fat Booty	Mos Def	
36	Freaky Tales	Too $hort	
37	Cave Bitch	Ice Cube	
38	Gangsta Bitch	Apache	
39	Just a Friend	Biz Markie	

Table 3. 3. List of the selected rap songs to form the new test set

CHAPTER 4

RESULTS AND DISCUSSION

This chapter reports all the results obtained during the model-building process. It begins with a baseline model that is built to select the best version of our dataset and for further comparison. Then, we discuss the performance of different models to select the best one, which is then optimized and implemented on new data for validation. This chapter ends with the error analysis to investigate the critical error and possible solutions for improving the model performance.

4.1 Building the Classification Models

4.1.1 Baseline model and preprocessing technique

A baseline LR model is iterated over the four preprocessing pipelines to select the best preprocessing technique/s. This step might play an important role in increasing the probability of obtaining higher results in the succeeding stages.

A general view from Figure 4.1 shows that ROC curve of LR model is performing well with all the pipeline where the difference among the pipelines is not descriptive. However, LR classifications are showing the highest ROC curve with raw data. Moreover, according to the data presented in Table 4.1, the LR model is more effective when data is not lemmatized, stemmed, lowercased, or does include stop words. Accuracy and sensitivity (recall) for the first two pipelines of the data are identical (79%), whereas raw data results in the highest AUC (87%) and the least FPR (20%). Therefore, we select the raw data for the proceeding analysis.

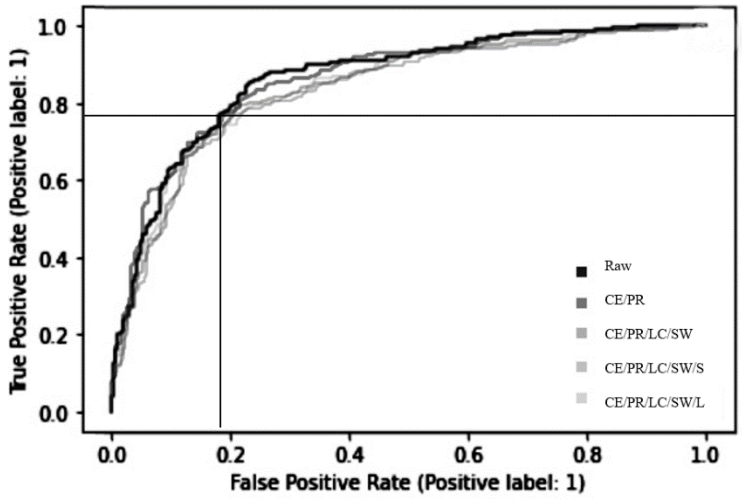

Figure 4. 1. Visualization of ROC curves of LR classifier using the different preprocessed versions of our dataset

Data	*AUC*	*TPR / Recall*	*FPR*	*Accuracy*
Raw	**0.87**	**0.79**	**0.20**	**0.79**
CE/PR	0.86	0.79	0.21	0.79
CE/PR/LC/SW	0.84	0.78	0.21	0.78
CE/PR/LC/SW/S	0.84	0.77	0.22	0.77
CE/PR/LC/SW/L	0.84	0.77	0.22	0.78

Table 4. 1. Results of the LR model on the different preprocessed versions of our dataset.

Subsequently, raw sentences are then split similarly to the baseline LR model and used for the application of the five selected classification models. Besides, the texts are vectorized using the BoW, word level TF-IDF, and N-gram TF-IDF to explore how models are performing with different feature extraction techniques. Table 4.2 shows that all classifiers are performing better when texts are vectorized using N-gram TF-IDF. The XGB classifier is, relatively, the least powerful among other models with the three feature extraction methods. The best performance is obtained using MNB and

SVM models. MNB is 1% more accurate, yet, both classifiers are effective and good in distinguishing between dehumanizing and non-dehumanizing sentences where MNB and SVM have a 90% AUC score. As displayed in Figure 4.2, the shaded area presents the range that is significant for our model, within this range, the TPR is maximized with the minimum FPR possible. The ROC curve of SVM is, generally, slightly higher than the MNB's curve but both classifiers show almost identical performance within the gray range. Therefore, we will optimize them both to find the best model that serves the objective of our study.

Model	*AUC*			*Accuracy*		
	BOW	*Word level TF-IDF*	*N-gram TF-IDF*	*BOW*	*Word level TF-IDF*	*N-gram TF-IDF*
NB	0.85	0.86	**0.9**	0.78	0.76	**0.83**
LR	0.86	0.86	0.89	0.78	0.78	0.81
SVM	0.86	0.88	**0.9**	0.79	0.79	**0.82**
RF	0.86	0.87	0.88	0.79	0.8	0.8
XGB	0.83	0.84	0.85	0.74	0.76	0.79

Table 4. 2. AUC scores resulting from the application of the five classification models using the version1 of the data.

4.1.2 Tuning the model

We Obtain good results with the initial SVM and MNB models. However, even better results might be achieved when optimizing the classifiers by tuning their hyperparameters to find the best combination of hyperparameters. Therefore, first, we select a dictionary that includes multiple values for each hyperparameter.

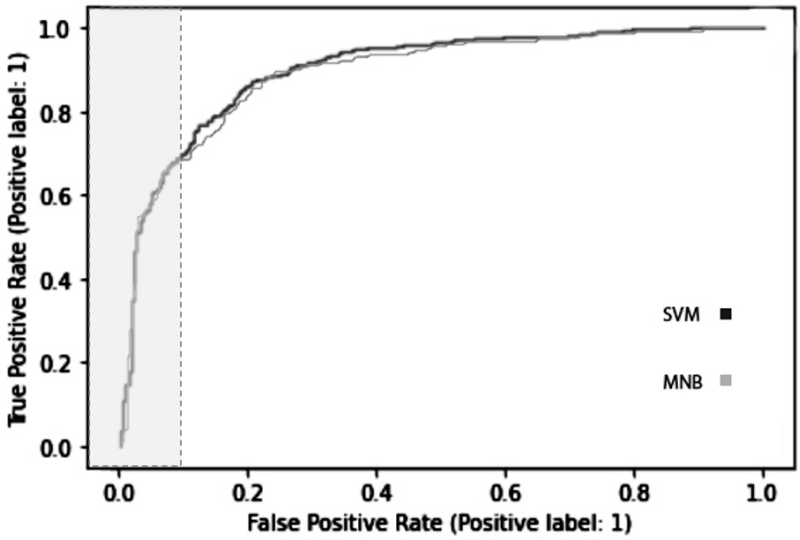

Figure 4. 2. Visualization of ROC curves of MNB and SVM classifiers using raw data

4.1.2.1 SVM hyperparameters tuning:

- The regularization parameter (C): its role is to control the trade-off between misclassifications and the margin width. For the optimization we select the five values [0.1, 1, 10, 100, 1000].
- Gamma: it is set before the training to give curvature weight for the decision boundary. We select the values [1, 0.1, 0.01, 0.001].
- Kernel: kernels help solve non-linear problems that have high dimensions without raising the complexity. We select three types: Polynomial, Gaussian RBF, and Sigmoid kernels.

By running the grid search method on our data to optimize the SVM classifier, the best parameters are shown, which are (C=10, gamma=0.1, kernel='sigmoid').

4.1.2.2 MNB hyperparameters tuning:

- Alpha: it is a hyperparameter that forms the model itself. It is usually given a value to resolve the issue of 0 probability and known as smoothing parameter. We select the values [1, 0.1, 0.01, 0.001, 0.0001, 0.00001]. The default is alpha=1 and this is the optimal value for our sample.
- Fit_prior : it tells the model whether to learn from prior class or not. We run both [True, False] and the best is fit_prior = False.

As shown in Figure 4.4, the overall performance of SVM and MNB gets better after the model tuning. The ROC curve of the optimal SVM model is higher than the initial SVM/MNB and the optimized MNB. More importantly, the FPR of the initial MNB/SVM starts to increase at almost 18% TPR, while the optimized MNB starts to have false positive classifications after achieving around 25% TPR. Whereas the curve of the optimized SVM is almost straight till it reaches a TPR of 0.4, which indicates that the model is effective in correctly classifying around 40% of the positive points with almost no false positives.

Digging deeper into the results to compare the optimized models, the tuned SVM has a 1% higher AUC score and is 2% more accurate than the tuned MNB, which indicates better overall classifications. The optimized SVM is less FPR by 3% than the optimized MNB, however, it has an 8% higher recall rate (Table 4.3). Our chosen model is the tuned SVM as it better serves the goal of this study in detecting women dehumanizing sentences.

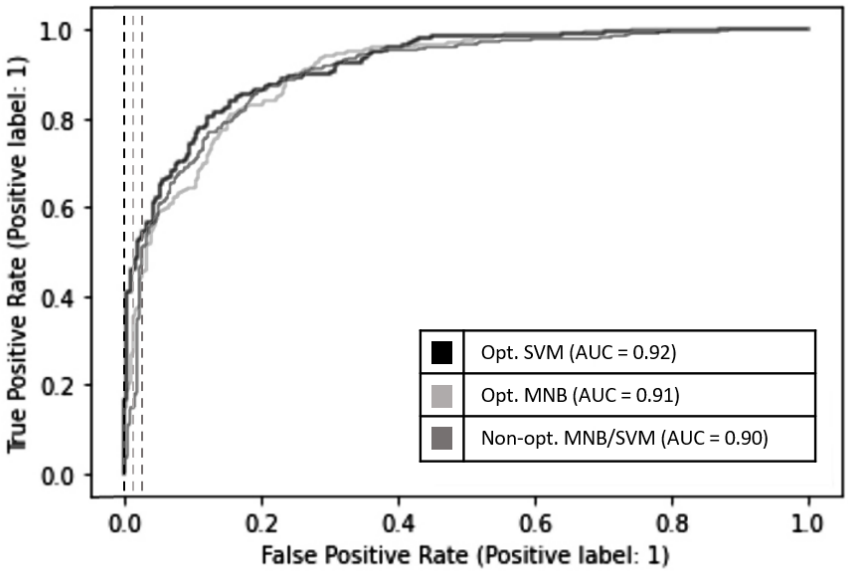

Figure 4. 3. Comparison between the ROC curve before and after the SVM/MNB optimization.

	AUC	*Accuracy*	*TPR*	*FPR*
Optimized MNB	0.91	0.82	0.77	0.13
Optimized SVM	**0.92**	**0.84**	**0.85**	0.16

Table 4. 3. Evaluating the performance of the SVM model before and after the optimization.

4.1.3 Threshold Selection

The ultimate objective of this study is to build a tool to accurately detect women dehumanizing sentences from an English text, which means, the goal is to maximize the recall of this model with the least error rate possible. Therefore, we need to define a classification threshold (i.e., the decision threshold). Any value higher than the specified threshold indicates “women dehumanization”; any value below indicates “non-dehumanization of women”.

From Figure 4.5, it is noticeable that TPR is increasing with the minimal value of FPR, since TPR reaches, approximately, 0.4 with almost zero FPR. After that, the curve starts to get steeper with the increase in sensitivity, which indicates a greater FPR value.

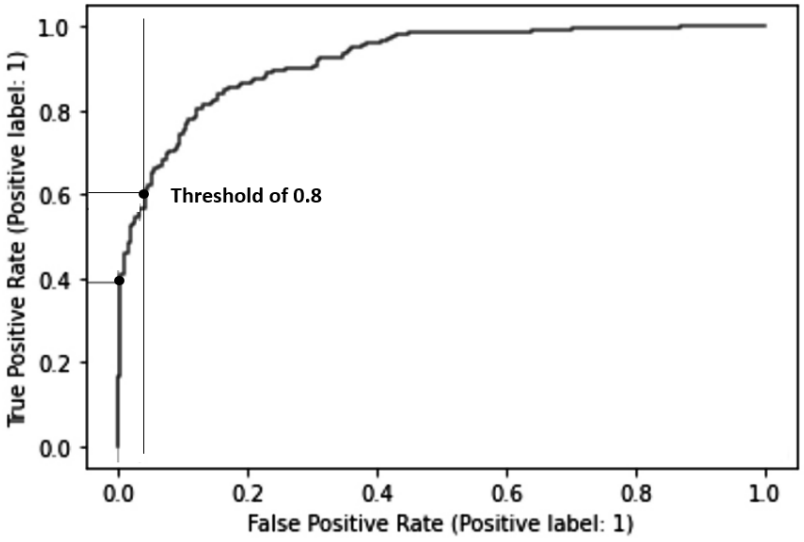

Figure 4. 4. ROC curve of the optimized SVM for threshold selection.

To find the optimal threshold for our objective, we further analyze the values of TPR, FPR, TNR, and precision for several thresholds. As reported in (Table 4.4), TPR starts to increase with no false predictions and 100% precision till the threshold of 0.95. The TRP continues to increase constantly till the threshold of 0.82. After this point, the change is only presented by a subtle increase in the false positives and a slight decrease in the precision. The TPR is consistent between 0.82 and 0.79, while precision continues in decreasing after 0.8. Therefore, we tolerate 5% FPR to get a 60% True detection rate we can obtain 60% of the True detection rate by setting 0.8 as the model's decision threshold.

Threshold	*FPR*	*TPR*	*Precision*
0.95	0	0.34	1
0.94	0.003	0.38	0.99
…	…	…	…
0.83	0.03	0.59	0.95
0.82	0.04	0.6	0.93
0.81	0.05	0.6	0.93
0.8	**0.05**	**0.6**	**0.93**
0.79	0.05	0.6	0.92

Table 4. 4. TPR, FPR, and precision at several classification thresholds

4.2 Performance Generalization on Unseen Data

An integral part of any machine learning system or tool is generalization on unseen or different kind of data. Testing the developed tool on real-world cases helps in ensuring the effectiveness of the final product and that it meets the target or purpose.

To validate our developed classifier, we test the performance of the trained model on new texts that are not seen by the model before and are more general and random in comparison to our built dataset. On that account, we collect lyrics on the selected rap songs as the new test set. Then, the developed model is run on every sentence of the songs lyrics to classify it according to the decision threshold. The results demonstrate that the developed model is powerful in detecting dehumanizing sentences. However, the predicted classifications need more investigation to evaluate the model performance more deeply when dealing with new data.

To analyze and validate our model performance, we check and relabel all sentences to plot ROC/AUC curve and compare the model's classifications and the true

classifications. As displayed in Figure 4.6, we plot the AUC/ROC curve of the model predictions with the true classifications (our labeling). Although the curve has a noticeable drop, figures show evidence that the model is still powerful since its true detection rate is 65% with 19% false positive rate.

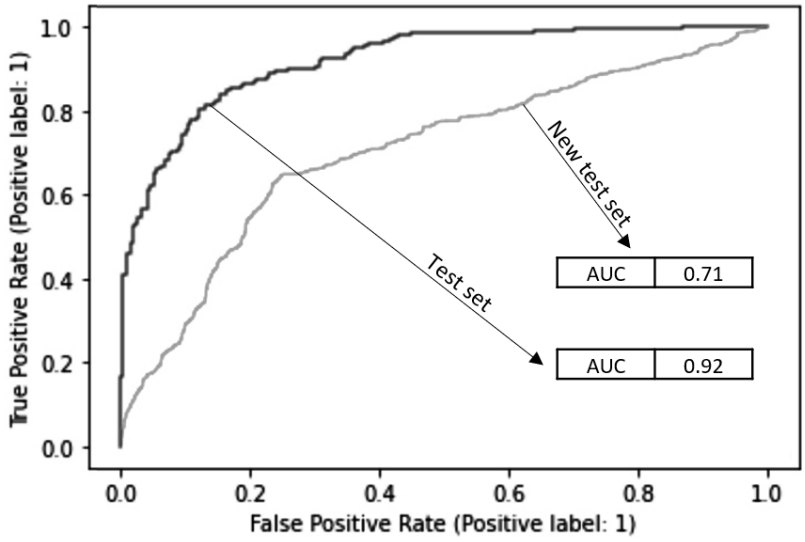

Figure 4. 3. A comparison between the test and validation AUC/ROC curves and TPR.

4.3 Error Analysis

The objective of this study is to automate the detection of women dehumanizing depiction in the text by creating a reliable and generalizable tool that has the least bias possible. Despite of that our model is tuned and evaluated based on several metrics, an in-depth analysis of the error is a key action that plays an important role in accomplishing the study's goal. On that ground, an in-depth review of the erroneous predictions of our model is performed according to the following steps:

4.3.1 Isolating sentences that are misclassified

A sample of the FPs which are represented by non-dehumanizing sentences that are incorrectly classified as dehumanizing and FNs that are represented by dehumanizing sentences that are incorrectly classified as non-dehumanizing, is presented in (Table 4.5) to ease the error diagnosis and identification.

1	FPs	Cars are crashin’ every night
2		She said her name was Donna
3		Two of the baddest in the history
4		She sold all his jewels she sold all his cars
5		And for the last 300 months
6		Come on over for a visit
7		Cause every time I turn on the TV
8		The first semester of the school year
9		It goes on and on and on, like that
10	FNs	To see my **baby doll**, I was happy to say
11		I neva love a broad cause I'm a **motherfucking** dog
12		Is **sugar** and **spice** the only thing that you made of?
13		Yous a tip drill, girl you a **tip drill**
14		Lookin’ like one of them **putty-cat dolls**
15		**Baby**, it's in yo nature (**meow**)
16		I get frequent flier mileage from my stewardess **chick**
17		I got this black **chick**, she don't know how to act

Table 4. 5. A sample from the model misclassifications.

4.3.2 Error identification

A comprehensive analysis of the model misclassifications leads us to the following considerations:

- **Abbreviations**: rap songs, usually, have lots of abbreviations that are not formal, in other words, lyrics are spelled as pronounced (e.g., crashin’

stands for crashing, neva stands for never, and yo stands for your). Slang language might be a reason behind the misclassifications.

- **Short sentences:** splitting the lyrics into sentences produces very short texts that are cut from a bigger context and are difficult to be understood alone.
- **Random negative examples:** songs have many sentences that are random and irrelevant to our dataset. To explain, our dataset contains positive and negative phrases that have the same patterns and include terms from source and target domains. However, examples 1 to 9 in Table 4.5 are falsely predicted to be positive, while they are general statements that do not have figurative language.
- **New source domain lexical items:** FNs have sentences with metaphors that are not included in our dataset, such as tip drill, baby doll, putty-cat doll, etc. (Table 4.5).
- Some metaphors are tricky and difficult to be found in a dehumanizing connotation using a specified pattern (e.g., baby and chick). Thus, our dataset might not have enough examples to train the model well.

4.3.3 Recommendations

After identifying possible hypotheses of the error, a few recommendations are given for further improvement of this detection tool:

1- To include random negative texts:

In the process of building our dataset, we include negative examples that have specific and limited patterns, however, random sentences might play an

important role in increasing the capability of the model in distinguishing between dehumanizing and non-dehumanizing examples.

2- Wider research on dehumanizing metaphors:

   Further research on figurative language is useful to find more dehumanizing and derogating metaphors.

3- To add sentences with slang terms.

4- Generalizing the model on different kind of datasets:

   The model might have better performance if tested on more formal language than rap songs.

4.4 Limitations and Future Work

This study has potential methodological limitations that, we think, are affecting the estimates and conclusion of our study.

4.4.1 Labeling collected sentences:

One of the limitations in our study is that only 3 out of 11 annotators are males. Having balanced number of males and females annotators might have different impact on the labeling of the collected sentences and, eventually, might result in different and more realistic models' predictions.

4.4.2 Lack of previous research works:

The lack of previous studies on this topic creates a need to develop the entire study almost from scratch, which makes it difficult to identify gaps. However,

discovering the limitations of this study is an important opportunity to ease the road for future studies in this area.

4.4.3 Vectorizing the text without considering the semantics of the sentences:

In our study, we use feature engineering techniques that might not consider the semantics of the sentences. The reason behind our choice is that the existing pre-trained word embeddings are probably biased. Training our models on biased semantics might affect the performance and, eventually, keep and convey these biased to future detection/mitigation tools.

Therefore, future research can be conducted in more realistic settings by considering the introduced limitations. In addition, future research could use more advanced techniques like BERT, which might raise the quality of the detection of the dehumanizing language against women. Moreover, further investigation on the dehumanizing language of both genders can be an interesting research topic.

CHAPTER 5
CONCLUSION AND FUTURE WORK

In summary, this paper discusses the importance of shedding the light on the dehumanizing language that is used against women by manifesting its negative impacts on the individual level, as well as, on the whole community. We, also, have shown the need to start building tools to detect and mitigate the use of this language. Therefore, as the detection of dehumanization from the text is a key step to making mitigation possible, our study is perhaps one of the first attempts to use machine learning algorithms to automatically detect women dehumanizing texts. Moreover, to our knowledge, this research work presents the first labeled dataset on women dehumanizing and derogating metaphorical expressions, which may facilitate future studies in this area.

Training several classification algorithms on the dataset that contains a wide range of metaphors, we find that a support vector machine classifier can detect women dehumanizing sentences with a 65% true detection rate with 19% false positive rate. The implication of the present findings is that the work on mitigating this language is doable at any point in time as a females dehumanizing statements can be detected and isolated using our developed model. Moreover, this model can be implemented to evaluate social media content, songs, movie scripts, and other texts to eliminate the spreading of this derogating language.

This study provides a good starting point for future discussions and further research on the mitigation of women degradation in the texts. Additionally, the provided dataset could be highly helpful for deeper investigations on this topic. Furthermore, a

number of recommendations for future research are given to improve the performance of the proposed model and raise the detection rate.

Despite the limitations of this study, our results demonstrate the effectiveness of our model in detecting the dehumanizing presentation of women in English texts, which might be the key component in future attempts to mitigate this type of sexism in the text.

BIBLIOGRAPHY

Allbritton, D. W. (1995). When Metaphors Function as Schemas: Some Cognitive Effects of Conceptual Metaphors. *Metaphor and Symbolic Activity*, *10*(1), 33–46. https://doi.org/10.1207/s15327868ms1001_4

Aslan, G. (2015). A metaphoric analysis regarding gender perceptions of preservice teachers. *Egitim ve Bilim*, *40*(181), 363–384. https://doi.org/10.15390/EB.2015.2930

Aslan, G. (2021). European Journal of Education Studies TEACHERS ' PERCEPTIONS OF GENDER : A METAPHORICAL ANALYSIS OF MALE AND FEMALE STUDENTS. *European Journal of Education Studies*, *8*(2), 362–383. https://doi.org/10.46827/ejes.v8i2.3586

Baider, F. H., & Gesuato, S. (2003). Masculinist Metaphors , Feminist research. *The Online Journal Metaphorik. De*, *5*, 6–25.

Bock, J., & Burkley, M. (2019). On the Prowl: Examining the Impact of Men-as-Predators and Women-as-Prey Metaphors on Attitudes that Perpetuate Sexual Violence. *Sex Roles*, *80*(5–6), 262–276. https://doi.org/10.1007/s11199-018-0929-1

Boroditsky, L. (2011). How Language Shapes Thought. *Scientific American, a Division of Nature America, Inc.*, *304*(2), 62–65.

Breiman, L. (2001). Random Forests. *Machine Learning*, 5–32. https://doi.org/10.1007/978-3-030-62008-0_35

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system.

Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, *13-17-August-2016*, 785–794. https://doi.org/10.1145/2939672.2939785

Cuddy, A. J. C., Fiske, S. T., & Glick, P. (2004). When professionals become mothers, warmth doesn't cut the ice. *Journal of Social Issues*, *60*(4), 701–718. https://doi.org/10.1111/j.0022-4537.2004.00381.x

Doughman, J. and, Khreich, W. and, El Gharib, M. and, Wiss, M. and, & Berjawi, Z. (2021). *Gender Bias in Text : Origin , Taxonomy , and Implications*. 1–10.

Egger, R. (2021). Applied Data Science in Tourism: Interdisciplinary Approaches, Methodologies, and Applications. In *Zeitschrift für Tourismuswissenschaft* (Vol. 13, Issue 2). https://doi.org/10.1515/tw-2021-0018

ElSherief, M., Kulkarni, V., Nguyen, D., Wang, W. Y., & Belding, E. (2018). Hate lingo: A target-based linguistic analysis of hate speech in social media. *12th International AAAI Conference on Web and Social Media, ICWSM 2018*, *ICWSM*, 42–51.

Erickson, T. (2002). Some problems with the notion of context-aware computing. *Communications of the ACM*, *45*(2), 102–104. https://doi.org/10.1145/503124.503154

Eyheramendy, S., D. Lewis, D., & Madigan, D. (2013). On the Naive Bayes Model for Text Categorization. In *Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics*. https://www.researchgate.net/publication/269107473_What_is_governance/link/548173090cf22525dcb61443/download%0Ahttp://www.econ.upf.edu/~reynal/Civil

wars_12December2010.pdf%0Ahttps://think-asia.org/handle/11540/8282%0Ahttps://www.jstor.org/stable/41857625

Goldenberg, J., Heflick, N., Vaes, J., Motyl, M., & Greenberg, J. (2009). Of mice and men, and objectified women: A terror management account of infrahumanization. *Group Processes and Intergroup Relations*, *12*(6), 763–776. https://doi.org/10.1177/1368430209340569

Harris, Z. S. (1954). Distributional Structure. WORD, *10*(2–3), 146–162. https://doi.org/10.1080/00437956.1954.11659520

Haslam, N. (2006). Dehumanization: An integrative review. *Personality and Social Psychology Review*, *10*(3), 252–264. https://doi.org/10.1207/s15327957pspr1003_4

Heflick, N. A., & Goldenberg, J. L. (2014). Seeing Eye to Body: The Literal Objectification of Women. *Current Directions in Psychological Science*, *23*(3), 225–229. https://doi.org/10.1177/0963721414531599

Hilbe, J. M. (2009). *Logistic Regression Models* (Vol. 15, Issue 2). Chapman and Hall/CRC.

Hines, C. (1999). Rebaking the Pie. *Evolution*.

Hosmer, D. W., & Lemeshow, S. (2013). Applied Logistic Regression, Thirs Edition. *Hoboken, NJ: Wiley-Interscience*, 1–30. https://doi.org/10.1002/0471722146.ch1

Jill C., S. (2011). Logistic regression: A brief primer. *Academic Emergency Medicine*, *18*(10). https://doi.org/10.1111/j.1553-2712.2011.01185.x

Kang, Hye-Min; Shaydullina, A. (2015). *Gender Construction in Stereotype-based*

Metaphors : Women as Desserts and as Animals. *August*. https://doi.org/10.13140/RG.2.1.4038.4161

Kuhn, M., & Johnson, K. (2013). Applied Predictive Modeling with Applications in R. In *Springer* (Vol. 26). http://appliedpredictivemodeling.com/s/Applied_Predictive_Modeling_in_R.pdf

Lakoff, G., & Johnson, M. (1987). Metaphors We Live By. In *Language, Thought, and Culture* (Vol. 35, Issue 2).

Li, C. (2019). Metaphors and Dehumanization Ideology: A critical analysis of the multimodal representation of women in advertising. *Chinese Semiotic Studies*, *15*(3), 349–377. https://doi.org/10.1515/css-2019-0021

López Maestre, M. D. (2020). Gender, Ideology and Conceptual Metaphors: Women and the Source Domain of the Hunt. *Complutense Journal of English Studies*, *28*, 191–206. https://doi.org/10.5209/cjes.68355

López Rodríguez, I. (2009). Of women, bitches, chickens and vixens: animal metaphors for women in English and Spanish. *Cultura, Lenguaje y Representación = Culture, Language and Representation: Revista de Estudios Culturales de La Universitat Jaume I = Cultural Studies Journal of Universitat Jaume I*, *7*, 77–100.

Lowd, D., & Domingos, P. (2005). *Naive Bayes Models for Probability Estimation*.

Mendelsohn, J., Tsvetkov, Y., & Jurafsky, D. (2020). A Framework for the Computational Linguistic Analysis of Dehumanization. *Frontiers in Artificial Intelligence*, *3*(August), 1–24. https://doi.org/10.3389/frai.2020.00055

Méndez, J. R., Iglesias, E. L., Fdez-Riverola, F., Díaz, F., & Corchado, J. M. (2006).

Lecture Notes in Artificial Intelligence: Preface. In *In Conference of the Spanish Association for Artificial Intelligence: Vol. 5180 LNAI*.

Menegatti, M., & Rubini, M. (2017). Gender Bias and Sexism in Language Introduction : Linguistic Processes and the Reproduction of Gender Bias. *Oxford Research Encyclopedia of Communication*, *September 2017*, 1–22.

Morris, K. L., Goldenberg, J., & Boyd, P. (2018). Women as Animals, Women as Objects: Evidence for Two Forms of Objectification. *Personality and Social Psychology Bulletin*, *44*(9), 1302–1314. https://doi.org/10.1177/0146167218765739

Morris, K. L., & Goldenberg, J. L. (2015). Women, objects, and animals: Differentiating between sex- and beauty-based objectification. *Revue Internationale de Psychologie Sociale*, *28*(1), 15–38.

Ortony, A. (1975). Why Metaphors Are Necessary and Not Just Nice. *Educational Theory*, *25*(1), 45–53. https://doi.org/10.1111/j.1741-5446.1975.tb00666.x

Porter, M. F. (2006). An algorithm for suffix stripping. *Program: Electronic Library and Information Systems*, *40*(3), 211–218. https://doi.org/10.1108/00330330610681286

Quilumba, F. L., Lee, W. J., Huang, H., Wang, D. Y., & Szabados, R. (2014). An overview of AMI data preprocessing to enhance the performance of load forecasting. *2014 IEEE Industry Application Society Annual Meeting, IAS 2014*, 1–7. https://doi.org/10.1109/IAS.2014.6978369

Recasens, M., Danescu-Niculescu-mizil, C., & Jurafsky, D. (2013). Linguistic models

for analyzing and detecting biased language. *ACL 2013 - 51st Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*, *1*, 1650–1659.

Reynolds, C., & Haslam, N. (2011). Evidence for an association between women and nature: An analysis of media images and mental representations. *Ecopsychology*, *3*(1), 59–64. https://doi.org/10.1089/eco.2010.0014

Rodríguez, I. L. (2007a). *Are Women Really Sweet? An Analysis Of The Women As Dessert Metaphor In The Engish And Spanish Written Press*. 179–195.

Rodríguez, I. L. (2007). The representation of women in teenage and women's magazines: recurring metaphors in English. *The Representation of Women in Teenage and Women's Magazines: Recurring Metaphors in English*, *15*(15), 15–42. https://doi.org/10.5209/rev_EIUC.2007.v15.8553

Rudman, L. A., & Mescher, K. (2012). Of Animals and Objects: Men's Implicit Dehumanization of Women and Likelihood of Sexual Aggression. *Personality and Social Psychology Bulletin*, *38*(6), 734–746. https://doi.org/10.1177/0146167212436401

Salton, G., Buckley, C. (1986). Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, *24*(5). https://doi.org/10.1163/187631286X00251

Sap, M., Gabriel, S., Qin, L., Jurafsky, D., Smith, N. A., & Choi, Y. (2019). *Social Bias Frames: Reasoning about Social and Power Implications of Language*. 5477–5490. https://doi.org/10.18653/v1/2020.acl-main.486

Schulder, M., & Hovy, E. (2015). *Metaphor Detection through Term Relevance*. *June*, 18–26. https://doi.org/10.3115/v1/w14-2303

Stefanowitsch, A. (2008). Corpus-based approaches to metaphor and metonymy. *Corpus-Based Approaches to Metaphor and Metonymy*, *January 2006*, 1–16. https://doi.org/10.1515/9783110199895.1

Tarkela, M. (2016). ”Being All BITCH BITCH BITCH Pms Pms BITCH BITCH BITCH” Conceptual Metaphors Describing Women on Seventeen.com and Cosmopolitan.com. *Proceedings of the 3rd Workshop on Gender Bias in Natural Language Processing*, 34--44.

Thibodeau, P. H., & Boroditsky, L. (2011). Metaphors we think with: The role of metaphor in reasoning. *PLoS ONE*, *6*(2). https://doi.org/10.1371/journal.pone.0016782

Tipler, C. N., & Ruscher, J. B. (2017). Dehumanizing representations of women: the shaping of hostile sexist attitudes through animalistic metaphors. *Journal of Gender Studies*, *28*(1), 109–118. https://doi.org/10.1080/09589236.2017.1411790

Tong, S., & Koller, D. (2001). Support Vector Machine Active Learning with Applications to Text Classification. *Journal of Machine Learning Research*, 45–66. https://doi.org/10.1353/aq.0.0077

Tourangeau, R., & Sternberg, R. J. (1982). Understanding and appreciating metaphors. In *Cognition* (Vol. 11, Issue 3, pp. 203–244). https://doi.org/10.1016/0010-0277(82)90016-6

Umera-Okeke, N. (2012). Linguistic sexism: An overview of the English language in

everyday discourse. *AFRREV LALIGENS: An International Journal of Language, Literature and Gender Studies*, *1*(1), 1–17.

Vujković, V., & Vuković-Stamatović, M. (2021). “What a kitty!”: women’s physical appearance and animal metaphors in montenegro. *Slovo*, *34*(1), 1–22. https://doi.org/10.14324/111.444.0954-6839.1239

Zhang, Y. (2012). Support vector machine classification algorithm and its application. *International Conference on Communications in Computer and Information Science*, *308 CCIS*(PART 2), 179–186. https://doi.org/10.1007/978-3-642-34041-3_27

Zhang, Y. T., Gong, L., & Wang, Y. C. (2005). Improved TF-IDF approach for text classification. *Journal of Zhejiang University: Science*, *6 A*(1), 49–55. https://doi.org/10.1631/jzus.2005.A0049